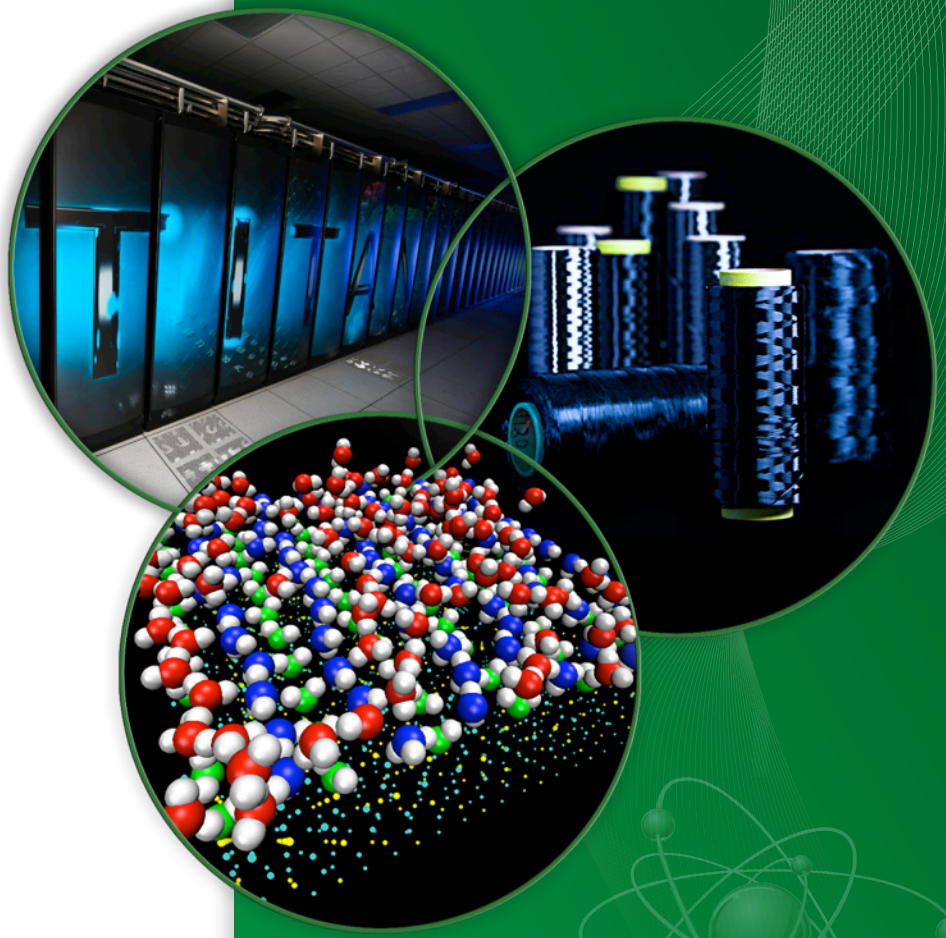
ORNL OLCF Facilities Plans



Jack Wells

Director of Science

Oak Ridge Leadership Computing Facility

Oak Ridge National Laboratory

HEP-ASCR Requirements Workshop

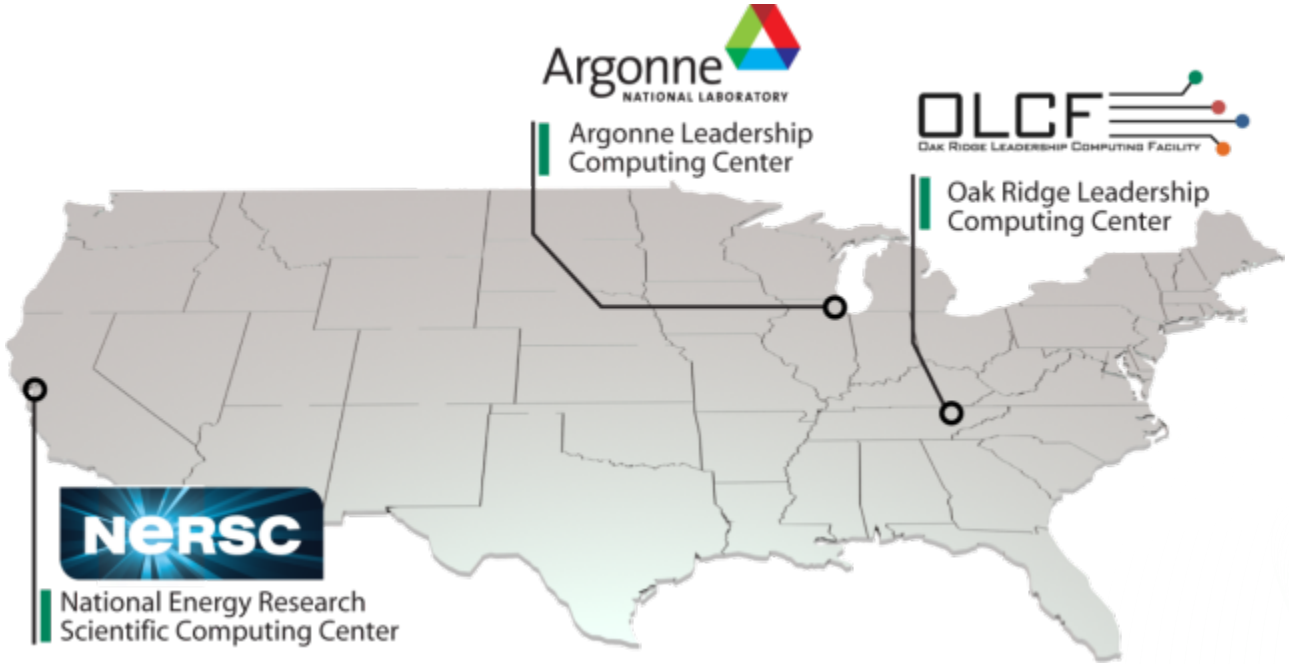
Bethesda

10 June 2015

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

DOE's Office of Science Computation User Facilities



- DOE is leader in open High-Performance Computing
- Provide the world's most powerful computational tools for open science
- Access is free to researchers who publish
- Boost US competitiveness
- Attract the best and brightest researchers

NERSC
Edison is 2.57 PF

ALCF
Mira is 10 PF

OLCF
Titan is 27 PF

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

What is the Leadership Computing Facility (LCF)?

- Collaborative DOE Office of Science user-facility program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).

OAK RIDGE National Laboratory

Argonne NATIONAL LABORATORY

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

What is CORAL (Partnership for 2017 System)

- CORAL is a Collaboration of Oak Ridge, Argonne, and Lawrence Livermore Labs to acquire three systems for delivery in 2017.
- DOE's Office of Science (DOE/SC) and National Nuclear Security Administration (NNSA) signed an MOU agreeing to collaborate on HPC research and acquisitions
- Collaboration grouping of DOE labs was done based on common acquisition timings. Collaboration is a win-win for all parties.
  - It reduces the number of RFPs vendors have to respond to
  - It improves the number and quality of proposals
  - It allows pooling of R&D funds
  - It strengthens the alliance between SC/NNSA on road to exascale
  - It encourages sharing technical expertise between Labs

Accelerating Future DOE Leadership Systems ("CORAL")

OAK RIDGE National Laboratory

"Summit" System

Lawrence Livermore National Laboratory

"Sierra" System

5X – 10X Higher Application Performance

IBM POWER CPUs, NVIDIA Tesla GPUs, Mellanox EDR 100Gb/s InfiniBand

Paving The Road to Exascale Performance

2017 OLCF Leadership System

Hybrid CPU/GPU architecture

SUMMIT
OAK RIDGE NATIONAL LABORATORY

OpenPOWER™

Vendor: IBM (Prime) / NVIDIA™ / Mellanox Technologies®

At least 5X Titan's Application Performance

Approximately 3,400 nodes, each with:

- Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta architecture
- CPUs and GPUs completely connected with high speed NVLink
- Large coherent memory: over 512 GB (HBM + DDR4)
  - all directly addressable from the CPUs and GPUs
- An additional 800 GB of NVRAM, which can be configured as either a burst buffer or as extended memory
- over 40 TF peak performance

Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect

IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.

How does Summit compare to Titan?

Feature	Summit	Titan
Application Performance	5-10x Titan	Baseline
Number of Nodes	~3,400	18,688
Node performance	> 40 TF	1.4 TF
Memory per Node	>512 GB (HBM + DDR4)	38GB (GDDR5+DDR3)
NVRAM per Node	800 GB	0
Node Interconnect	NVLink (5-12x PCIe 3)	PCIe 2
System Interconnect (node injection bandwidth)	Dual Rail EDR-IB (23 GB/s)	Gemini (6.4 GB/s)
Interconnect Topology	Non-blocking Fat Tree	3D Torus
Processors	IBM POWER9 NVIDIA Volta™	AMD Opteron™ NVIDIA Kepler™
File System	120 PB, 1 TB/s, GPFS™	32 PB, 1 TB/s, Lustre®
Peak power consumption	10 MW	9 MW

Two Tracks for Future Large Systems

Tianhe-2 (NUDT): TH-IVB-FEP
Intel Xeon E5-2692 12 C 2.2 GHz
TH Express-2
Intel Xeon Phi 31S1P

Titan (Cray): Cray XK7
AMD Opteron 6274 16C 2.2 GHz
Cray Gemini
NVIDIA K20x

Sequoia (IBM): BlueGene/Q
Power BQC 16C 1.6 GHz

K computer (Fujitsu)
SPARC64 VIIIfx 2.0 GHz
Tofu

Mira (IBM): BlueGene/Q
PowerPC A2 16C 1.6 GHz

Piz Daint (Cray): Cray XC30
Intel Xeon E5-2670 8C 2.6 GHz
Cray Aries
NVIDIA K20x

Edison (Cray): Cray XC30
Intel Xeon E%-2695v2 12C 2.4 GHz
Aries

Many Core

- 10's of thousands of nodes with millions of cores
- Homogeneous cores
- Multiple levels of memory – on package, DDR, and non-volatile
- Unlike prior generations, future products are likely to be self hosted

Hybrid Multi-Core

- CPU / GPU Hybrid systems
- Likely to have multiple CPUs and GPUs per node
- Small number of very fat nodes
- Expect data movement issues to be much easier than previous systems – coherent shared memory within a node
- Multiple levels of memory – on package, DDR, and non-volatile

Cori at NERSC

- Self-hosted many-core system
- Intel/Cray
- 9300 single-socket nodes
- Intel® Xeon Phi™ Knights Landing (KNL)
- 16GB HBM, 64-128 GB DDR4
- Target delivery date: June, 2016

Summit at OLCF

- Hybrid CPU/GPU system
- IBM/NVIDIA
- 3400 multi-socket nodes
- POWER9/Volta
- More than 512 GB coherent memory per node
- Target delivery date: 2017

ALCF-3 at ALCF

- 3rd Generation Intel Xeon Phi (Knights Hill (KNH)
- > 50,000 compute nodes
- Target delivery date: 2018

ASCR Computing Upgrades At a Glance

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrade
Name Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori 2016**	**Summit 2017-2018**	**Aurora 2018-2019**
System peak (PF)	2.6	27	10	> 30	150	180
Peak Power (MW)	2	9	4.8	< 3.7	10	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	> 7 PB DRAM and persistent memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 15 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Hill many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	150 PB 1 TB/s Lustre®

CORAL Acquisitions

Center for Accelerated Application Readiness: Summit

OLCF-4 issued a call for proposals in FY2015 for application development partnerships between community developers, OLCF staff and the OLCF Vendor Center of Excellence.

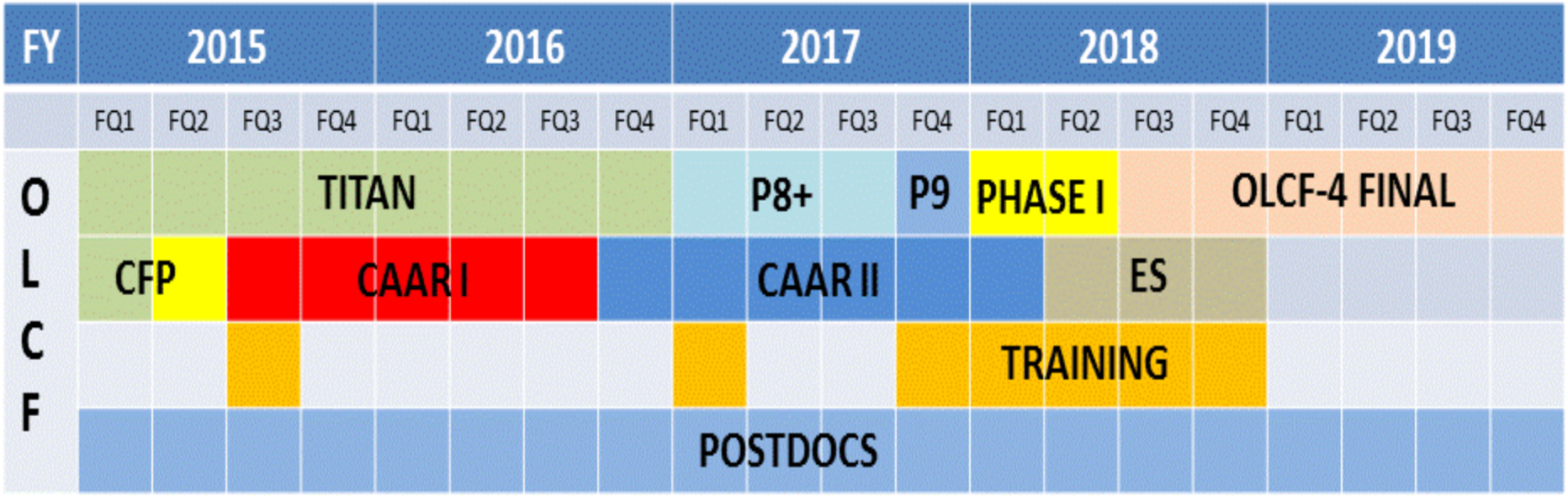
Center for Accelerated Application Readiness (CAAR)

- Performance analysis of community applications
- Technical plan for code restructuring and optimization
- Deployment on OLCF-4

New Application Readiness Activities CAAR

Application	Domain	Principal Investigator	Institution
ACME (N)	*Climate Science*	David Bader	Lawrence Livermore National Laboratory
DIRAC	*Relativistic Chemistry*	Lucas Visscher	Free University of Amsterdam
FLASH	*Astrophysics*	Bronson Messer	Oak Ridge National Laboratory
GTC (NE)	*Plasma Physics*	Zhihong Lin	University of California – Irvine
HACC(N)	*Cosmology*	Salman Habib	Argonne National Laboratory
LSDALTON	*Chemistry*	Poul Jørgensen	Aarhus University
NAMD (NE)	*Biophysics*	Klaus Schulten	University of Illinois – Urbana Champaign
NUCCOR	*Nuclear Physics*	Gaute Hagen	Oak Ridge National Laboratory
NWCHEM (N)	*Chemistry*	Karol Kowalski	Pacific Northwest National Laboratory
QMCPACK	*Materials Science*	Paul Kent	Oak Ridge National Laboratory
RAPTOR	*Engineering*	Joseph Oefelein	Sandia National Laboratory
SPECFEM	*Seismic Science*	Jeroen Tromp	Princeton University
XGC (N)	*Plasma Physics*	CS Chang	Princeton Plasma Physics Laboratory

CAAR Timeline



1. November 2014: Call for CAAR applications
2. February 20, 2015: CAAR proposal deadline
3. March 2015: Selection of CAAR application teams
4. **April 2015: CAAR application training workshop**
5. April 2015: CAAR application teams start
6. June 2016: CAAR project review
7. October 2017: Call for Early Science projects
8. November 2017: Selection Early Science projects
9. January 2018: Early Science projects start
10. October 2018: Early Science project ends

CAAR in Preparation of Summit

Application Developer Team involvement

- Knowledge of the application
- Work on application in development “moving target”
- Optimizations included in application release

Early Science Project

- Demonstration of application on real problems at scale
- Shake-down on the new system hardware and software
- Large-scale science project is strong incentive to participate

Vendor technical support through the IBM/NVIDIA Center of Excellence is crucial

- Programming environment often not mature
- Best source of information on new hardware features

Access to multiple resources, including early hardware

Joint training activities

Portability is a critical concern

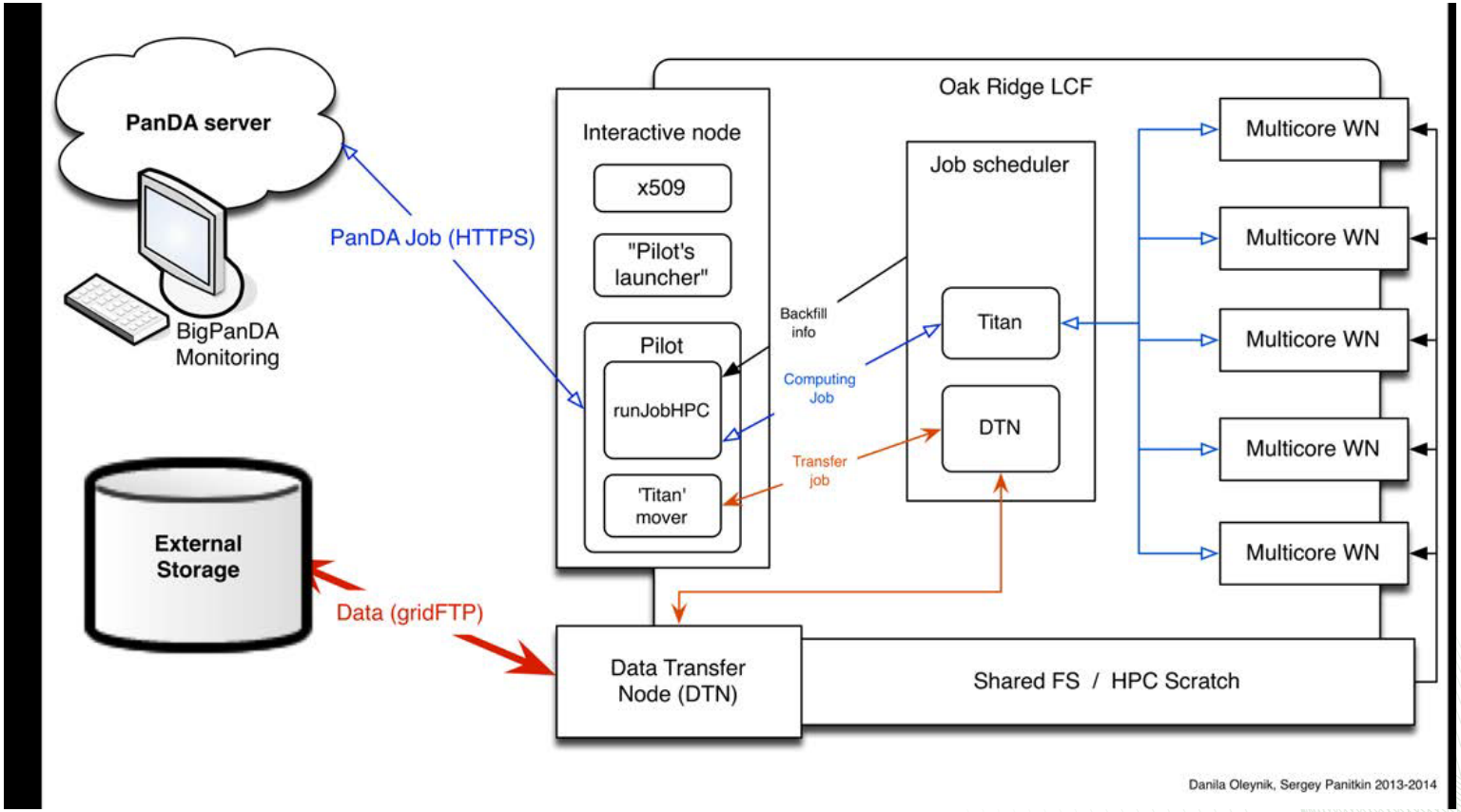
PanDA Tool Provides Titan with Next-Gen Workflow for Big Data

- Researchers with the ATLAS experiment in Europe have been integrating its scheduling and analysis tool, PanDa, with Titan.
- Global PanDA workflow includes 1.8 million jobs each day distributed among 100 or so computing centers spread across the globe.
- PanDA's ability to efficiently match available computing time with high-priority tasks holds great promise for Titan.
- Team developers redesigned parts of the PanDA system on Titan responsible for job submission on remote sites and gave PanDA new capability to collect information about unused worker nodes on Titan.
- Deployment of the tool could lead to a higher utilization of available hours on Titan.
  - Three day test in July 2014 increased Titan utilization by 2.5%.

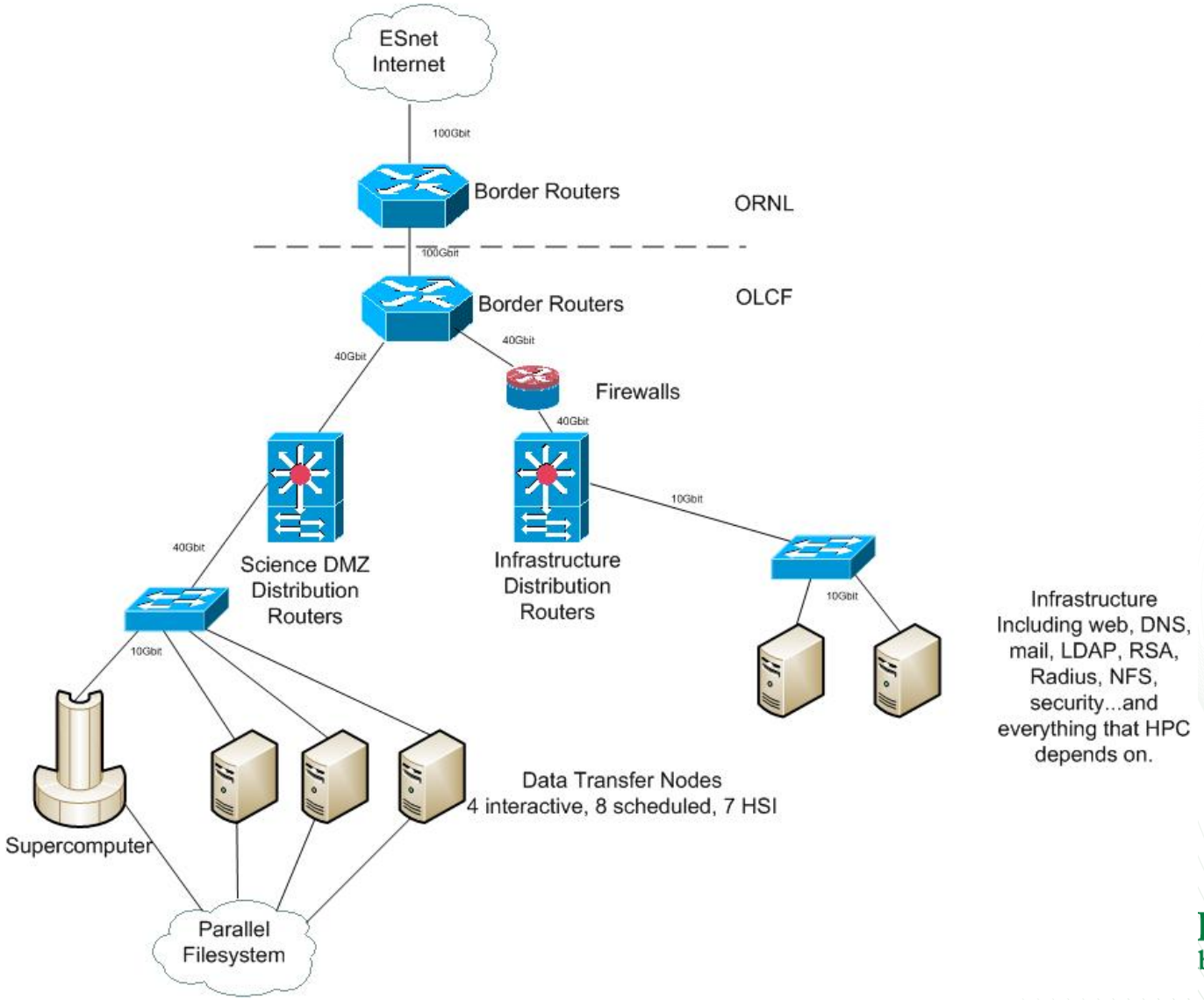
PanDA manages all of ATLAS's data tasks from a server located at CERN, the European Organization for Nuclear Research.

PanDA architecture for Titan

- Pilot(s) executes on HPC interactive node
- Pilot interact with local job scheduler to manage job
- Data, produced on HPC automatically moves to external storage



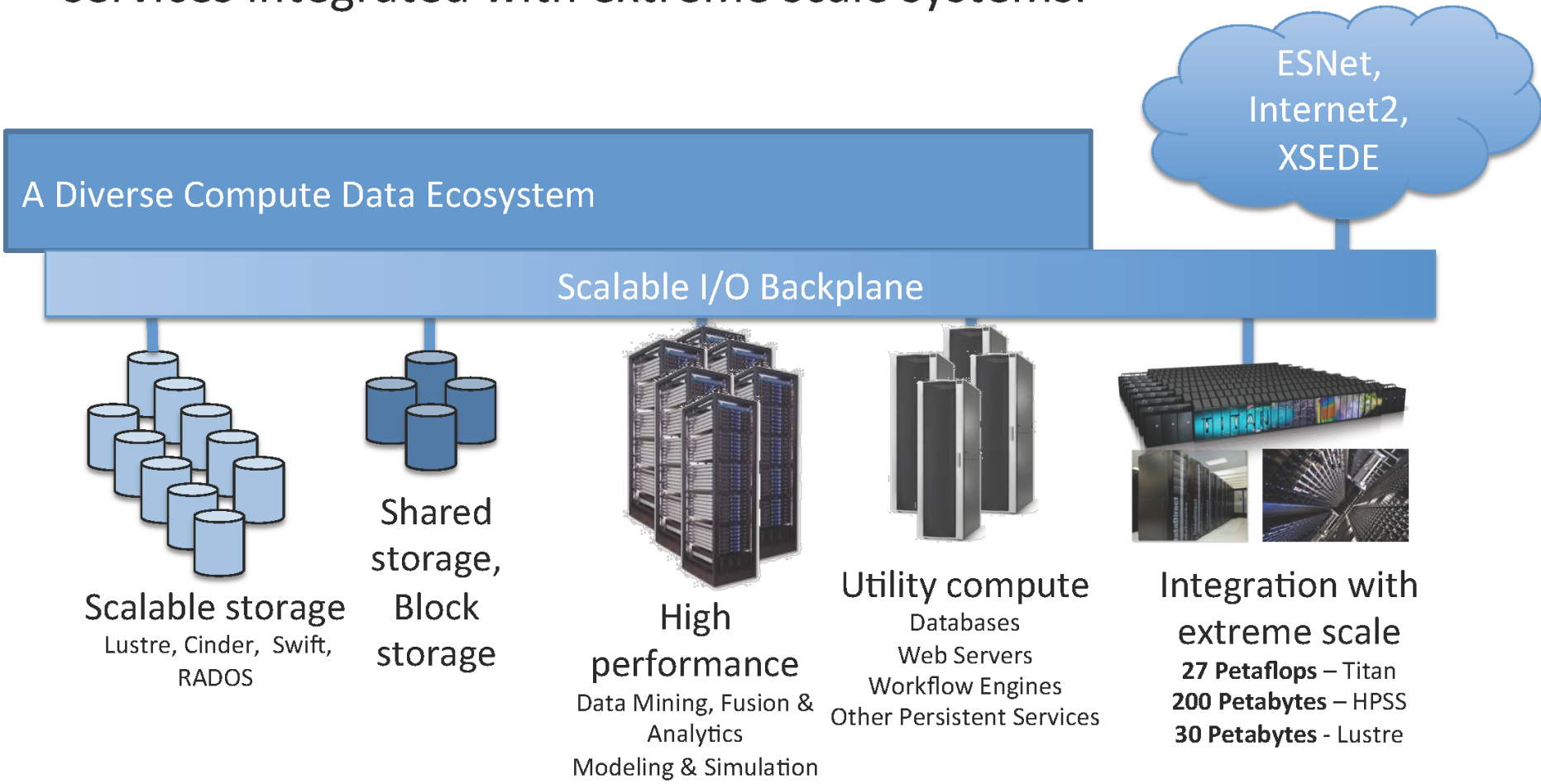
OLCF & ESNet are implementing the Science DMZ to enable high-performance access to ESNet WAN



CADES

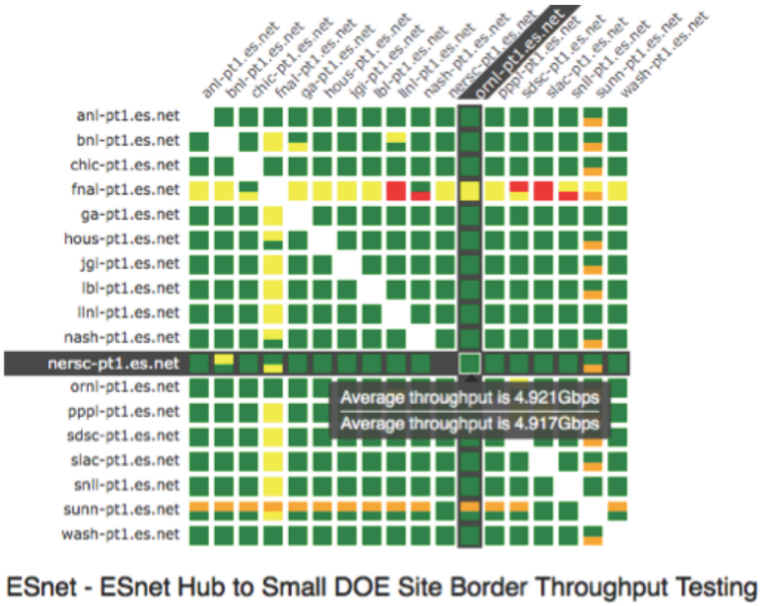
Compute & Data Environment for Science

- A diverse ecosystem of compute and data infrastructure and services integrated with extreme scale systems.



ALICE-USA Project Plan:
New ALICE T2 facility at Oak Ridge National Laboratory

- ORNL CADES Facility:
  - Compute And Data Environment for Science
- LBNL NERSC & ORNL CADES
  - Both with Scientific Computing strength
  - High-bandwidth, monitored uplink to ESnet
  - Proximity to DOE HPC Resources with strategic alignment to O^2 project
    - Oak Ridge Leadership Computing Facility (OLCF)
    - NERSC DOE SC Flagship facility



ESnet - ESnet Hub to Small DOE Site Border Throughput Testing

- 2014 Project Proposal:
  - Establish new distributed ALICE T2 facility
    - Continue deployment & operation at LBNL NERSC
    - Establish new T2 site early 2015 at **ORNL CADES**
    - Transition operations from LLNL to ORNL in 2015
- Project Proposal Reviewed, June 2014
  - Proposal endorsed with modest recommendations